Integrating association rules mined from health-care data with ontological information for automated knowledge generation

John Heritage[1], Sharon McDonald[2] and Ken McGarry[1]

[1] School of Pharmacy and Pharmaceutical Sciences,
Faculty of Health Sciences and Wellbeing,
University of Sunderland, City Campus, UK.
ken.mcgarry@sunderland.ac.uk
[2] Faculty of Computing,
University of Sunderland, St Peters Campus, UK.

Abstract. Association rule mining can be combined with complex network theory to automatically create a knowledge base that reveals how certain drugs cause side-effects on patients when they interact with other drugs taken by the patient when they have two or more diseases. The drugs will interact with on-target and off-target proteins often in an unpredictable way. A computational approach is necessary to be able to unravel the complex relationships between disease comorbidities. We built statistical models from the publicly available FAERS dataset to reveal interesting and potentially harmful drug combinations based on side-effects and relationships between co-morbid diseases. This information is very useful to medical practitioners to tailor patient prescriptions for optimal therapy.

Keywords: comorbidity, side-effect, association rules, support, confidence, pharmaco-epidemiology

1 Introduction

As people age and suffer from several illnesses, they will require more medications. When individuals start taking several medications the chances that the drugs they take will interact in harmful ways will increase. Drug-to-drug interactions are difficult to predict as there are so many confounding factors at work - people vary in their genetic predisposition and thus response to treatment, age, gender, and environmental factors all play a role. Although every drug undergoes rigorous safety trials during its development, these are conducted on participants using only the drug being investigated, it is impossible to conduct the trial any other way. Our knowledge of drug-to-drug interactions, side-effects and disease comorbidity is derived from healthcare record systems and these are now starting to receive increased attention as a way of improving public health and drug safety.

Collecting data on drug side-effects and carefully analyzing it can reveal much about how drugs are acting in the body and should assist doctors tailor drug prescriptions for their patients [12,11]. This is made possible by identifying shared biological pathways

through similar side-effects. The USA and UK have online systems such as the FAERS [16] and *Yellow Card* [5] databases in place for medical professionals to report incidents when patients experience an adverse drug reaction (ADR). Unfortunately, there is a great deal of noise present in these databases and in fact potentially the majority of cases may be anecdotal and unreliable. For example a patient, in the early stages of drug treatment may present themselves at the doctors complaining of headaches, dizziness and feelings of nausea. Some symptoms may not be listed on the drug information sheet and there is a chance it is not a result of taking the drug, perhaps the patient has an additional undiagnosed condition or had taken medicine for flu. However, the value of *big data* patient records comes from the luxury of being able to discard the poor quality, noisy cases and to keep only a fraction [18]. Powerful statistical models need only a few hundred high quality records to perform reliable comparisons that can unravel the complex interactions between drug regimens, patient variability and random chance.

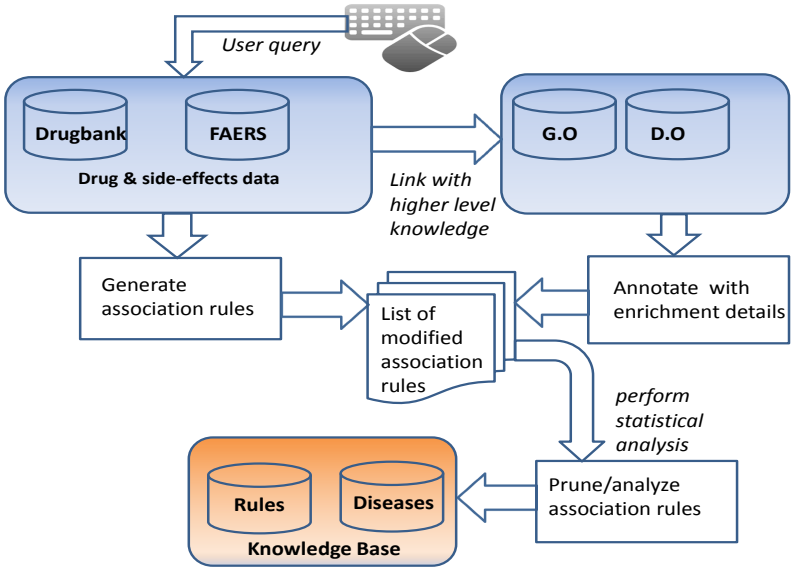

Fig. 1: Overview of system operation, showing database sources, data flow and statistical analysis. The user query initiates a search of the various databases resulting in a knowledge base and ruleset related to the disease of interest.

Referring to figure 1 the system is intended to be used by healthcare specialists wishing to test a hypothesis relating to the drugs their patients are currently receiving for a particular disease, their patient may already have a second medical condition and be taking medicine for this. The practitioner would like to query the system to see if

anything is known about potential drug conflicts. Clearly, some drug-to-drug interactions are already well known and there would be no need to use this system, but for suspected combinations of drugs the system would be useful.

The process is initiated by a user query containing the disease(s) of interest which accesses the various databases, the results of which are used to build statistical models to assess drugs and to build the knowledge base. We briefly describe the key methods and data.

Association rule mining and frequent item set analysis originated in market basket analysis 20 years ago whereby trends and patterns in consumer purchasing activities could be identified and used to increase profits [1]. Several companies use association rules to influence customers to purchase items, most notably the Amazon recommendation facility which is based on an individual's purchasing history and also the purchasing history of others with similar tastes and preferences [3].

Association rule mining is now starting to receive attention in the bioinformatics field where the majority of the data are binary or categorical in nature. So far, the bulk of association rules research in bioinformatics appears to be oriented towards mining Electronic Health Records (EHR) and Health Information Technology (HIT) on patient clinical information [23] and little attention focused on drug-to-side-effects [22]. In this paper association rules are used to uncover relationships between drugs, their side-effects and co-interacting protein targets. However, it should be noted that the item sets can only show the commonality of items as they appear in the database, however by using association rules and a statistical measure we can imply strong correlation between the associated items.

The databases used include Drugbank for the chemical properties and protein targets of the candidate drugs and SIDER4 for a comprehensive list of drug side-effects. Furthermore, we used the Gene Ontology (GO) for providing details and characteristics of specific proteins and products [2], the Disease Ontology (DO) which relates the various diseases into a taxonomy [17]. KEGG is a store of biological pathways and provides more information on the normal biological process of the cell and how they can be affected by drugs [14].

The remainder of this paper is structured as follows; section two describes the system architecture in terms of flow of information, the sources of data and the computational techniques we use; section three describes the results; section four is the discussion; finally section five presents the conclusions and future work.

2 Methods

2.1 Programming environment

We implemented the system using the R language with the RStudio programming environment. R is primarily a statistical data analysis package but is gaining popularity for various scientific programming applications and is very extendable using packages written by other researchers [15]. We used the following R packages: aRules [7], GOSim [24]. Our R code and data files are freely available to all researchers on GitHub for download: https://github.com/kenmcgarry/UKCI2017-AR

2.2 Databases, ontologies and pre-processing

The Gene Ontology (GO), KEGG and Disease Ontology (DO) are used to annotate the proteins and drug targets with additional information useful for a deeper interpretation of the biological processes and structures [8,17]. The DO database contains knowledge on 8,043 inherited, developmental and acquired human diseases. Through enrichment analysis, the R package DOSim is able to explore the biological meaning of related genes in terms of structure, function and hierarchy. The concepts in DOSim are organized into a directed acyclic graph (DAG) similar to a tree structure, the concepts are linked through various relationships. The lower the term or concept is positioned in the hierarchy then the more specific the term is, higher-up terms describe higher level or more abstract concepts. The ontologies are used to tag the association rules with biological meaning, in the sense they provide the medical users some indication of the pathways that are affected by the drugs, and how such pathways may lead to specific side-effects being presented by the patient. Incidentally, this kind of information is useful to drug companies as they are now desperate to reuse existing drugs (repurpose) for potentially very different diseases to the ones they were originally designed to treat [10,12,20].

The FDA (Food and Drug Administration) provide a freely available database called the adverse event reporting system (AERS). Online reporting of FDA AERS occurs on a quarterly basis, and began in January of 2004. Legacy records are available through the National Technical Information Service on compact disc (on a fee basis) or downloaded electronically. A quarterly AERS report contains several subsets of information:

- Demographics (DEMO); basic patient information.
- Drug types (DRUG); a list of drugs administered to patients.
- Indications (INDI); why they were given the drug initially.
- Outcomes (OUTC); the end result, e.g. hospitalisation, death.
- Reactions (REAC); side effect(s) experienced.
- Reporting sources (RPSR); where the information originated from.
- Therapy types (THER); how and when the drug was administered.

These seven subsets are linked using the primary key ID we can identify any patient with the drugs, side-effects and diseases they suffer from. During the course of a year a given patient may have more than one ADR and can appear several times. The data is all text based and stored in flat files which we saved in (CSV) format to enable access by our R programming environment.

As it pertains to the prediction of associations, all such reporting systems are inherently acute in nature; the information they contain has already been filtered based upon a close temporal association observed by those reporting to the databases. A prime example of how analysis of such databases may be lacking in predictive powers would be the development of cancers; a drug may induce the formation of a cancer, but the significant degree of latency between exposure and the formation of an observable symptom will likely be too great for some possible causal agents to be considered worth entering into the database.

2.3 Association rule mining

In a database, when considering the occurrence of one item with another in the same record (or transaction) represents an association. The frequency with which these items appear together overall in the database may represent some important relationship or trend. Several techniques are available such as the $Apriori$ algorithm that can extract rules which highlight these occurrences and their frequencies. Formally we can define the following: where $\mathcal{I} = \{I_1, I_2, \ldots, I_m\}$ are a set of items. Let $\mathcal{D}$ be a collection of transactions in a database, where each transaction t has a unique identifier and contains a set of items such that $t \subseteq \mathcal{I}$. A set of items is called an *itemset*, and an itemset with k items is called a *k-itemset*.

A number of statistics are available to rank and order the association rules such as the $support$, $confidence$ and $lift$ of a rule. The *support* of an itemset x in $\mathcal{D}$, denoted as $\sigma(x/\mathcal{D})$, is the ratio of the number of transactions (in $\mathcal{D}$) containing x to the total number of transactions in $\mathcal{D}$.

An *association rule* is an expression $x \Rightarrow y$, where $x, y \subseteq \mathcal{I}$ and $x \cap y = \emptyset$. The *confidence* of $x \Rightarrow y$ is the ratio of $\sigma(x \cup y/\mathcal{D})$ to $\sigma(x/\mathcal{D})$. There are several association rule interestingness measures available should the number of extracted rules be large in number. The measures will select only those rules that have a certain statistical strength and confidence, and thus prune down the number rules to a manageable size.

Perhaps more helpfully we can say from a shopping database:
$\{bread, cheese\}\ x \Rightarrow y\ \{butter\}$

If a customer buys bread and cheese, then they are also likely to buy butter.

Lift is analogous to the relative response of patients and is of primary interest in identifying novel adverse events as this metric accounts for the high frequency of some consequences; e.g. nausea is a frequent consequence of many drugs combinations and if confidence alone were used to order associations this consequence, and others like, it would reduce the signal to noise ratio of the survey by appearing in every other record.

2.4 Related work

The MOAL (Multi Ontology At All Levels) system of Manda *et al* [9] is dedicated to extracting meaningful patterns and relationships from the Gene Ontology. When presented a gene product/disease, MOAL will generate association rules across all three sub-databases and use these to annotate the new gene products. Manda *et al* have also developed their own interestingness metrics to evaluate and assess the discovered rules: *MOConfidence* and *MOSupport* derive the necessary information from a cross-ontology platform and select the most informative rules, thus pruning any superfluous information. [13] Uncovering disease-disease relationships through the incomplete human interactome. The DIseAse MOdule Detection (DIAMOnD) algorithm by Ghiassian used a systematic analysis of connectivity patterns of disease proteins in the human interactome [6]. Tatonetti analyzed drug interactions from adverse-event reports where they discovered interaction between paroxetine and pravastatin that increased blood glucose levels, thus warning doctors about this combination [19]. The work of Cai et al is similar to ours [4], they also use association rules but frame these in the context of Bayesian networks in an attempt to explain causality. We use complex network theory to frame our rules.

3 Results

The drugs were first filtered to find the most frequent one hundred in recognition of fact that these have a propensity to occupy the majority of drug entries (approx 47% of the dataset) and that there is an apparent window of filtering that occurs around this region in which the ratio of side effect per drug entry reaches a maximum before again decreasing; the quantity of frequently occurring drugs filtered out can be easily adjusted within the code to examine the extents of the filters impact. These highly frequent drugs were then mapped back to the original drug list to mark patient entries containing them. The primaryid of the patients consuming the more frequent drugs was extracted and used as a master filter to remove all entries corresponding to those particular patient cases from both the drug and side effect lists. This ID always us to uniquely identity the patient and to link all known ADR's this person has suffered along with drugs they take, the diseases they suffer from. Since we are using one years worth of data (2016), we also miss the patient's history of previous ADR's and their drugs. Although this is not our objective in this work, we realize the importance of these databases to track trends in disease development over time, the drugs used and perhaps discarded through ADR's.

Table 1: Data mining parameters for Apriori algorithm on the FAERS dataset

Parameter	Value
Number of patients remaining in dataset	364,368
Number of drug and side effect observations in filtered dataset	19,790
Number of unique drug and side effect observations in filtered dataset	15,743
Apriori minimum support threshold	0.000005
Apriori minimum confidence	0.000001
Apriori minimum rule length	3
Apriori maximum rule length	3
Number of association rules initially generated	718,662
Number of association rules when constrained for drug antecedent and side effect consequent	778,820
Rules with at least ten observations present	368

Rules are then generated in the form of an antecedent (left hand side, LHS) $=>$consequent (right hand side, RHS) relationship. Table 1 shows the initial setup for the association rule algorithm and the early set of results. Finally, the rules generated were filtered to remove those with less than ten observations, those with none unique side effects (as many similar drug combinations produce similar side effects) and sorted by lift criterion. The results were validated by comparing to Stockleys interaction checker available via British National Formulary.

In figure 2 the frequency of observations (drugs and side effects) per patient case after filtering off cases containing most frequent hundred drugs and subsequently cases containing the hundred most frequent side effects. The majority of patients have around ten or less observations in their records. 364,368 unique patients under examination, displaying 15,743 observations (drugs or side effects) at an array density of 0.02%.

Referring to figure 3 illustrates the distribution of rules found on the basis of their lift over the number of times they were observed. When data mining, it can be useful to have

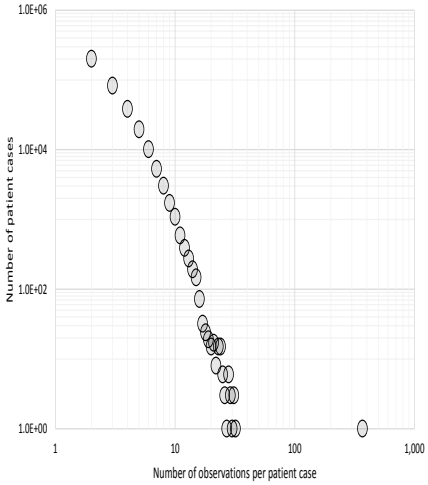

Fig. 2: number of side-effects per patient case

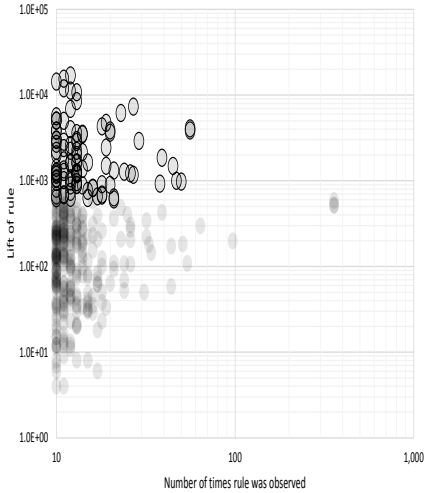

Fig. 3: number of rules per patient case

the axes include all possible values as it is one of the few occasions available to visually survey for anomalies in transformations; the datasets themselves are far too large to survey by eye. Rules that will be investigated further are encircled with a perimeter.

In figure 5 the highest lift associations made ten or more times. The rule will be stated along with its number of observations and lift. It will be proceeded by an explanation of the medical terminology, then any and all interactions listed by Stockleys and the rules predictability.

Rather than sift through the extracted rule-base, a health-care practitioner who is interested in a specific disease or indication and wishes to control how the rules are extracted they need to get the correct text name or UMLS code for their query. The system at the moment is highly sensitive to case/spelling and words used. Future work will address these shortcomings and make the system more robust to user errors and differences in nomenclature.

This produced six rules with low support but high confidence, the lift criteria was substantive. These are shown in table 3, the PU entry refers to *Product used for unknown indication* which is a frequently occurring item for the majority of drugs and implies that they are used for off-the-label diseases. Conversely, taking the problem the other way around what can association rules tell us about the comorbidities a patient can suffer from if they do develop Atrial fibrillation? Table 4 displays these rules.

The next stage was to calculate and integrate semantic similarity between our comorbid diseases identified by the association rules and to pull out any valid connections between them. Whilst the scope of the work described in this paper is beyond formal knowledge representation of any gene-disease associations using Gene Ontology (GO), we have Disease Ontology (DO) which provides a consistent description of gene prod-

	rules	support	confidence	lift
1	{Low density lipoprotein increased} => {Cardiovascular event prophylaxis}	0.20	82.64	397.15
2	{Cardiovascular event prophylaxis} => {Low density lipoprotein increased}	0.20	55.02	397.15
3	{Low density lipoprotein increased} => {Blood cholesterol increased}	0.18	76.80	90.20
4	{Blood cholesterol increased} => {Low density lipoprotein increased}	0.18	12.50	90.20
5	{Myelofibrosis} => {Product used for unknown indication}	0.20	46.14	1.36
6	{Cardiac failure} => {Product used for unknown indication}	0.18	40.84	1.21
7	{Epilepsy} => {Product used for unknown indication}	0.22	30.18	0.89
8	{Chronic myeloid leukaemia} => {Product used for unknown indication}	0.22	24.03	0.71
9	{Schizophrenia} => {Product used for unknown indication}	0.23	28.40	0.84
10	{Cardiovascular event prophylaxis} => {Blood cholesterol increased}	0.19	53.49	62.83
11	{Blood cholesterol increased} => {Cardiovascular event prophylaxis}	0.19	13.07	62.83
12	{Gait disturbance} => {Multiple sclerosis}	0.29	42.76	9.39
13	{Multiple sclerosis} => {Gait disturbance}	0.29	3.69	9.39
14	{Gait disturbance} => {Product used for unknown indication}	0.51	75.65	2.24
15	{Prostate cancer} => {Product used for unknown indication}	0.26	22.21	0.66
16	{Pulmonary hypertension} => {Product used for unknown indication}	0.36	34.84	1.03
17	{Bipolar disorder} => {Product used for unknown indication}	0.19	28.08	0.83
18	{Seizure} => {Product used for unknown indication}	0.22	29.41	0.87
19	{Deep vein thrombosis} => {Product used for unknown indication}	0.24	33.42	0.99
20	{Ankylosing spondylitis} => {Product used for unknown indication}	0.19	13.57	0.40

Table 2: Top scoring rules for all indications based on lift criteria

	lhs	rhs	support	confidence	lift
6	{Cerebrovascular accident prophylaxis,PU,Thrombosis prophylaxis}	{Atrial fibrillation}	0.00	0.99	47.30
3	{Cerebrovascular accident prophylaxis,Thrombosis prophylaxis}	{Atrial fibrillation}	0.01	0.98	47.03
5	{Cerebrovascular accident prophylaxis,PU}	{Atrial fibrillation}	0.01	0.80	38.31
4	{PU,Thrombosis prophylaxis}	{Atrial fibrillation}	0.00	0.77	36.83
2	{Cerebrovascular accident prophylaxis}	{Atrial fibrillation}	0.01	0.76	36.35
1	{Thrombosis prophylaxis}	{Atrial fibrillation}	0.01	0.63	30.04

Table 3: Comorbidities associated with Atrial Fibrillation generated six rules. It supports the question : What problems are patients likely to suffer from before developing Atrial Fibrillation? The left hand side (LHS) contains the antecedent and the right hand side (RHS) contains the consequent. Where: $*$PU (Product used for unknown indication)

ucts with disease perspectives, and is essential for supporting functional genomics in disease context.

The comorbidities are identified using their local identifiers in DO. A number of measures can be used to rank semantic similarity, here we used the *Wang* criterion as it reflects the biological plausibility better than other measures because of the way semantic similarity of the DO terms are calculated, using both the locations of the terms in the DO graph and their relations with their ancestor terms [21].

$$Wang(A,B) = \frac{\sum_{t \in T_A \cap T_B} S_A(t) + S_B(t)}{SV(A) + SV(B)} \quad (1)$$

For the Wang equation, where $SA(t)$ represents the S-value of DO term t related to term A and $SB(t)$ is the S-value of DO term t related to term B.

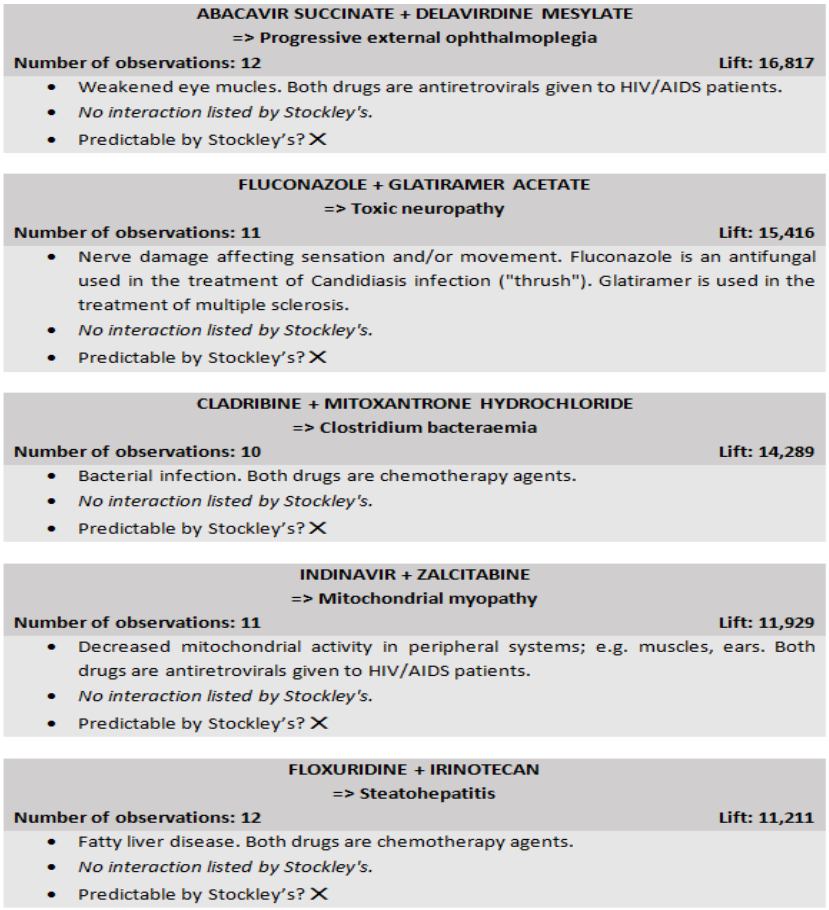

Fig. 4: Top five ADR's ranked by lift criteria

3.1 Ontology integration

Individually, each method of data analysis provides important information in a specific area. However, further value of comes from the integration of these disparate sources of knowledge in a principled way. We use a variation of the Jaccard similarity coefficient to integrate the many sources of heterogenous information into a coherent entity for decision making.

$$Score = \text{disease}_{ij} + \text{drugs}_{ij} + \text{sideeffects}_{ij} + \text{DO-similarity}_{ij} + \text{GO-similarity}_{ij}$$

$$\text{Score index}_{ij} = \frac{|F(D_i) \cap F(D_j)|}{|F(D_i) \cup F(D_j)|}$$

	lhs	rhs	support	confidence	lift
2	{Atrial fibrillation}	{Cerebrovascular accident prophylaxis}	0.01	0.60	36.35
4	{Atrial fibrillation}	{Product used for unknown indication}	0.01	0.39	1.14
1	{Atrial fibrillation}	{Thrombosis prophylaxis}	0.01	0.37	30.04
3	{Atrial fibrillation}	{Hypertension}	0.00	0.11	3.18

Table 4: Comorbidities associated with Atrial Fibrillation generated four rules. It supports the question what problems will patients suffer from if they develop Atrial Fibrillation? This time the left hand side (LHS) contains the antecedent and the right hand side (RHS) contains the consequent. Where: ∗PU (Product used for unknown indication)

Where $F(D_j)$ are the features of interest, such as disease, side-effects, the drugs $F(D_i)$. The rules are then reevaluated using these scores and re-ranked. Implementing the equation produces a matrix with the diagonal containing the Jaccard score for the combination of association rule and the attached ontological terms.

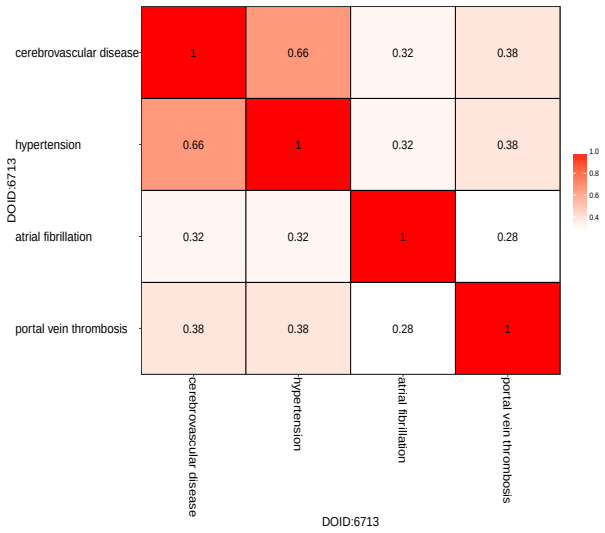

Fig. 5: Correlation matrix for the disease similarity

The strongest correlation is between hypertension and cerebrovascular disease (.66), hardly a novel discovery but does reveal that this method is useful for integrating association rules with the semantic similarity of any disease.

4 Discussion

We processed and analyzed 162,744 patient entries, 24,641 unique drugs and 8,025 unique side effects were surveyed over four hours in a referenced study, using an ap-

proximately similar methodology, yielding 2,603 association rules at a minimum observation occurrence of 50. The results presented here began with 3,045,688 patients entries, 8,106 unique drugs, 16,248 unique side effects, was completed in approximately ten minutes and produced 78,820 association rules that were further refined to 368 at a minimum observation occurrence of just 2. An 18 fold increase in the number of patients that can be examined and 24 fold reduction in the time required has been achieved.

Some drug interactions were not found in Stockleys because the active ingredients used were not recognised; e.g. HYPROMELLOSE 2910 (4000 MPA.S) is not listed, but hypromellose is. Whilst no interactions were found for some drugs, the side effects for a drug alone could suggest the possible outcome. Alternatively, someone skilled or otherwise in the field of pharmacology could likely predict some of the outcomes based on medical experience. The approach on which this work was based employed the skills of a pharmacovigilance expert to assess the results for predictable interactions. Drug names were used verbatim here, and no attempt was made to further guess at the likelihood of a possible interaction outcome as, in practice, an algorithm lacking such worldly knowledge would be incapable of doing this without first having a reference for such predictable interactions, and it is the application of algorithms to medical records that is of interest.

There are interesting observations present; for instance, that the combination of antipsychotics results in bleeding and bruising, even though these drugs are thought to be highly specific to neuronal functionality. Assurance that the methodology followed is able to detect rare events is exemplified by ANENOCOUMAROL + LEVOFLOXACIN $\rightarrow$ International normalized ratio increased and EFAVIRENZ + ETONOGESTREL $\rightarrow$ Pregnancy with implant contraceptive; both rare occurrences per Stockleys and high lift results where obtained.

In terms of the association rule generator (Apriori) we found that the candidate generation could be extremely slow based on the number of elements in LHS (pairs, triplets, etc.). Furthermore, the candidate generation process could generate duplicates depending on the implementation. The counting method iterates through all of the transactions each time.

During our analysis we found if a pathology is frequently encountered, such as heart attacks, a connection may not be drawn between the outcome and its possible causal agent being a particular drug or combination of them. As a result, it is likely that such chronic conditions are under-represented in the current works. Conversely, some entries may be over-reported as a result of media attention, accidental data entry, legal issues, a product being newer to the market.

The issues involved with the FAERS database have proven difficult to resolve, the manner in which the FDA list sequences of events for a particular patients entry has changed over time. In its totality, the database is neither entirely structured or unstructured in form; information not only moves location within the database, with some fields being deleted, created or translocated, the associated field headers also change. Demographics such as gender was originally listed as *gndr-cod* and subsequently as *sex*. The same field, like others, has also moved location within the database.

5 Conclusions

There may be better approaches for finding less frequent (more novel) patterns, as apriori itself is fundamentally intended for finding frequent patterns. Huge quantities of the results apriori generates from medical records are just very common, very well known side effects; warfarin + aspirin $\rightarrow$ bleeding is a common discovery in the literature. However, when searching for more novel observations, it is more appropriate to reverse that, starting at the lower frequencies and working up towards the frequent rules. We achieved something roughly similar with the software we developed by first filtering off all the patients consuming frequent drugs and experiencing frequent side effects before running the algorithm; around half of the all the drug entries correspond to just 10 drugs in the FDA records. Future work will involve a more effective highlighting of unique drug combinations which may be achieved by initially filtering on a combinatorial basis instead; only excluding drugs that form frequent combinations. This would be most rapidly achieved using the first pass (FP-growth) hierarchical tree algorithm.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: SIGMOD-93. pp. 207–216 (1993)
2. Ashburner, M.: Gene ontology: tool for the unification of biology. Nature Genetics 25, 25–29 (2000)
3. Brin, S., Motwani, R., Silverstein, C.: Beyond market baskets: generalizing association rules to correlations. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. pp. 265–276 (1997)
4. Cai, R., Liu, M., Hu, Y., Melton, B.L., Matheny, M.E., Xu, H., Duan, L., Waitman, L.R.: Identification of adverse drug-drug interactions through causal association rule discovery from spontaneous adverse event reports. Artificial Intelligence in Medicine 76, 7 – 15 (2017), http://www.sciencedirect.com/science/article/pii/S0933365716305437
5. Dunn, N., Mann., R.: Prescription-event and other forms of epidemiological monitoring of side-effects in the uk. Clinical and Experimental Allergy 29(3), 217–239 (1999)
6. Ghiassian, S., Menche, J., Barabasi, A.: A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS Computational Biology 11(4) (2015)
7. Hahsler, M., Chelluboina, S., Hornik, K., Buchta, C.: The arules R-package ecosystem: Analyzing interesting patterns from large transaction datasets. Journal of Machine Learning Research 12, 1977–1981 (2011)
8. Li, J., Gong, B., Chen, X., Liu, T., Wu, C., Zhang, F., Li, C., Li, X., Rao, S., Li, X.: Dosim: An R package for similarity between diseases based on disease ontology. BMC Bioinformatics 12(1), 266 (2011), http://www.biomedcentral.com/1471-2105/12/266
9. Manda, P., McCarthy, F., Bridges, S.: Interestingness measures and strategies for mining multi-ontology multi-level association rules from gene ontology annoations for the discovery of new GO relationships. Journal of Biomedical Informatics 46(5), 849–856 (2013)
10. McGarry, K.: Discovery of functional protein groups by clustering community links and integration of ontological knowledge. Expert Systems with Applications 40(13), 5101–5112 (2013)

11. McGarry, K., Emery, K., Varnakulasingam, V., McDonald, S., Ashton, M.: Complex network based computational techniques for edgetic modelling of mutations implicated with human diseases. In: The 16th UK Workshop on Computational Intelligence, UKCI-2016. pp. 89–105. Springer-Verlag, University of Lancaster, UK (7th-9th September 2016)
12. McGarry, K., Slater, N., Amaning, A.: Identifying candidate drugs for repositioning by graph based modeling techniques based on drug side-effects. In: The 15th UK Workshop on Computational Intelligence, UKCI-2015. University of Exeter, UK (7th-9th September 2015)
13. Menche, J., Sharma, A., Kitsak, M., Ghiassian, S., Vidal, M., Loscalzo, J., Barabasi, A.: Uncovering disease-disease relationships through the incomplete human interactome. Science 347(6224) (2015)
14. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M.: Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 27, 29–34 (1998)
15. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2015), https://www.R-project.org/
16. Rodriguez, E., Staffa, J., Graham, D.: The role of databases in drug postmarketing surveillance. Pharmacoepidemiol Drug Safety 10(5), 407–410 (2001)
17. Schriml, L., Arze, C., Nadendla, S., Chang, Y.W., Mazaitis, M., Felix, V., Feng, G., Kibbe, W.: Disease ontology: a backbone for disease semantic integration. Nucleic Acids Research 40, D940 – D946 (2012)
18. Tatonetti, N., Fernald, G., Altman, R.: A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports. J Am Med Inform Assoc 19(1), 79–85 (2012)
19. Tatonetti, N.P., Denny, J.C., Murphy, S.N., Fernald, G.H., Krishnan, G., Castro, V., Yue, P., Tsao, P.S., Kohane, I., Roden, D.M., Altman, R.B.: Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. Clinical pharmacology and therapeutics 90, 133–42 (2011)
20. Wang, F., Zhang, P., Cao, N., Hu, J., Sorrentino, R.: Exploring the associations between drug side-effects and therapeutic indications. Journal of Biomedical Informatics 51, 15–23 (2014)
21. Wang, J., Du, Z., Payattakool, R., Yu, P., Chen, C.: A new method to measure the semantic similarity of GO terms. Bioinformatics 23(10), 1274–1281 (2007)
22. Wright, A., Chen, E., Maloney, F.: An automated technique for identifying associations between medications, laboratory results and problems. Journal of Biomedical Informatics 43(6), 891–901 (2010)
23. Yang, J., Li, Z., Fan, X., Cheng, Y.: Drug disease association and drug-repositioning predictions in complex diseases using causal inference probabilistic matrix factorization. Journal of Chemical Information and Modeling 54(9), 2562–2569 (2014)
24. Yu, G., Yan, G., He, Q.: DOSE: an R/Bioconductor package for disease ontology semantic and enrichement analysis. Bioinformatics 31(4), 608–609 (2015)